GAO
Accountability * Integrity * Reliability

United States General Accounting Office
Washington, DC 20548

Accounting and Information
Management Division

B-286259.1

September 29, 2000

The Honorable Dan Miller
Chairman
The Honorable Carolyn B. Maloney
Ranking Minority Member
Subcommittee on the Census
Committee on Government Reform
House of Representatives

Subject: 2000 Census: Update on Data Capture Operations and System

As you know, earlier this year the U.S. Census Bureau adopted a two-phase approach to capturing household data for the 2000 decennial census. Under phase one, which the bureau terms first pass, only the data necessary to determine the reapportioning of seats in the House of Representatives, which the bureau calls 100 percent data, are captured. Under the second pass, the remaining data, which the bureau calls sample data, are captured.

To implement this two-pass approach, the bureau had to modify one of its key systems, the Data Capture System (DCS) 2000, so that during the first pass only the 100 percent data from the digitally imaged census forms (short and long) would be optically read (and keyed) and so that the long-form images could be written to a mass storage device. At the same time that it has been conducting first-pass operations, the bureau's DCS 2000 development contractor has been creating a modified version of the system so that during second-pass operations the stored long-form images can be retrieved and the sample data can be optically scanned and keyed.

At your request, we have been reviewing and periodically reporting to you on (1) the bureau's progress in performing first-pass data capture operations, including the performance of DCS 2000, and (2) the bureau's progress in modifying DCS 2000 to

perform planned second-pass data capture operations.[1] This letter provides our findings on both topics since we last reported to you in May 2000 testimony, and in light of our findings, concludes our work on data capture operations and DCS 2000.

In summary, we found that the bureau has made good progress on first-pass data capture operations. In particular, the bureau met its milestone for completing its processing of questionnaires that households mailed back, and it finished processing of questionnaires completed by enumerators ahead of schedule. In addition, DCS 2000 has exceeded its key performance goals, such as optical character recognition (OCR) accuracy, in support of first-pass operations. We similarly found that the bureau's contractor responsible for modifying DCS 2000 to support second-pass operations is proceeding slightly ahead of schedule and that the contractor is employing effective management controls in testing the modified system.

Background

The Constitution requires a decennial census in order to reapportion seats in the House of Representatives among the states. In addition, public and private decisionmakers use census data on population counts and social and economic characteristics for a variety of purposes (e.g., state and local redistricting; allocations of government funding; and planning and evaluation activities, such as site selection for new schools, market research, and evaluations of local labor markets). Also, the census is the only national source of detailed population statistics for small geographic areas, such as towns and school districts, and for population groups, such as Native Americans.

Since 1970, the bureau has used essentially the same methodology to capture census data. The bureau develops an address list of the nation's housing units and delivers census forms to those housing units, requesting that occupants mail back the completed forms. Most households are sent a short form to complete; however, some are asked to complete a long form.[2] The bureau then hires temporary census-takers, known as enumerators, by the hundreds of thousands to gather the requested data for each nonresponding housing unit. Captured data are then provided to

---

[1] *2000 Census: New Data Capture System Progress and Risks* (GAO/AIMD-00-61, February 4, 2000); *2000 Census: Status of Key Operations* (GAO/T-GGD/AIMD-00-91, February 15, 2000); *2000 Census: Update on Essential Operations* (GAO/T-GGD/AIMD-00-119, March 14, 2000); *2000 Census: Progress Report on the Mail Response Rate and Key Operations* (GAO/T-GGD/AIMD-00-136, April 5, 2000); and *2000 Census: Status of Nonresponse Follow-up and Key Operations* (GAO/T-GGD/AIMD-00-164, May 11, 2000).

[2] The census short form—three pages with eight questions—was delivered to approximately 83 percent of all housing units. It asked the respondent to provide information for up to six household members including information regarding name, age, sex, relationship, and race. The long form—40 pages with 53 questions—was delivered to approximately 17 percent of all housing units. For up to six household members, it asked the same questions as the short form as well as questions on social, economic, financial, and physical characteristics.

Census headquarters for further processing and tabulation and for generation o+f census products.

To conduct the 2000 census, the bureau is relying on 10 key systems. These systems enable the bureau to develop and maintain address lists, maps, and geographic reference files; collect census data through the Internet; scan and process household-completed paper forms; analyze census data; recruit and support temporary workers; facilitate follow-up surveys; and track costs and performance related to taking the census. DCS 2000 is one of these key systems. It performs the following high-level functions: checks in completed forms, creates digital images of the completed forms, optically reads the responses on the forms, converts these data into files, and transmits them to bureau headquarters for tabulation and analysis. Specific DCS 2000 subsystems are

- the *data verification and receipt subsystem*, which among other operations, (1) receives the paper census forms and prepares them for imaging and (2) identifies respondents so the bureau can identify and follow up with nonrespondents;
- the *scanning and imaging subsystem*, which creates an electronic image of the paper form;
- the *optical recognition subsystem*, which captures census data from the electronic form images;
- the *keying subsystem*, which is used to manually input data that cannot be satisfactorily read from the paper form; and
- the *data preparation function*, which formats data from the optical recognition and keying subsystems and then sends the data to Census Bureau headquarters.

DCS 2000 is located at four data capture centers (DCC) in Baltimore, Maryland; Pomona, California; Phoenix, Arizona; and Jeffersonville, Indiana. The Jeffersonville location is the bureau-operated National Processing Center while the other three sites are temporary facilities provided and operated by a bureau contractor. The bureau is acquiring, deploying, and maintaining DCS 2000 through a system development contractor.

Two-Phase Approach to Data Capture Operations: A Brief Description

As a result of DCS 2000 operational testing in late 1999, the bureau realized that it could not process census forms fast enough to meet its master schedule for completing Census 2000 and delivering apportionment counts to the Congress by December 31, 2000. To resolve this dilemma, the bureau adopted the two-pass approach to data capture operations, which required it to modify DCS 2000. During

the first pass—from March 6, 2000 until September 14, 2000—the DCCs captured only the data necessary to determine the apportionment counts. During the second pass—planned from August 28, 2000 to November 15, 2000—the DCCs are capturing the remaining data from long census questionnaires, which include the detailed social, economic, and housing information collected for a sample of households in the United States.[3]

To implement the two-pass data capture solution, two sets of changes—or releases—were required for DCS 2000. The first release, designed to support the first pass, was completed in early February. This work involved modifying DCS 2000 software to write the long-form images to a mass storage unit and to not present the sample data to keyers. The second release, designed to support the second pass, involves modifying the system to retrieve the images of the more than 23.5 million long-form questionnaires from the mass storage unit and present those requiring action to keyers, and then transmitting the resulting data to bureau headquarters. The bureau and its DCS 2000 development contractor expected to complete the second-release changes on July 30, 2000, then install and test the changes at the Baltimore DCC from July 31 through August 25. Changing DCS 2000 to the two-pass approach resulted in estimated cost increases of $33 million for additional system development, hardware, integration, testing, and support by the development contractor; and $12 million for the contractor that operates the DCCs to keep the centers operational longer than originally planned.

We initially reported on DCS 2000's status in February 2000, when we said that the bureau had made considerable progress on the system but that much work remained to be accomplished and little time remained to accomplish it before data capture operations were to begin in early March 2000.[4] Subsequently, we provided periodic testimony to your subcommittee that included updates on the bureau's progress in completing DCS 2000, the status of first-pass data capture operations, the operational performance of DCS 2000, and the progress being made in modifying DCS 2000 for second-pass data capture.[5] When we last testified before your subcommittee in May 2000, we reported that the bureau's DCCs were processing questionnaires at a rate that would meet their May 26 deadline for completing mailback questionnaire processing. We also testified that the DCS 2000 development contractor (1) was meeting its master-plan commitments for modifying DCS 2000 so that it could retrieve and process the long-form questionnaires and (2) had adopted an appropriate risk-based approach to modifying DCS 2000's hardware and software configurations.

---

[3]The conclusion of first-pass operations was phased across the four DCCs beginning with the Baltimore DCC on July 25 and concluding with the Jeffersonville DCC on September 14. Similarly, the initiation of second-pass operations is phased across the DCCs beginning with the Baltimore DCC on August 28 and concluding with the Jeffersonville DCC on October 10.

[4]GAO/AIMD-00-61, February 4, 2000.

[5]GAO/T-GGD/AIMD-00-91, February 15, 2000; GAO/T-GGD/AIMD-00-119, March 14, 2000; GAO/T-GGD/AIMD-00-136, April 5, 2000; and GAO/T-GGD/AIMD-00-164, May 11, 2000.

Nevertheless, we noted that much remained to be accomplished and plan execution was the key to success of the second pass.

Objectives, Scope, and Methodology

Our objectives were to determine (1) the bureau's progress in performing first-pass data capture operations, including the performance of DCS 2000, and (2) the bureau's progress in modifying DCS 2000 to perform planned second-pass data capture operations.

To determine the bureau's progress in performing data capture operations, we reviewed operations plans, schedules, models, and goals (e.g., goals for OCR accuracy, Key From Image (KFI) accuracy, and keying rate). We then analyzed progress and performance information from the bureau's data capture performance database, reviewed bureau and contractor status reports, and compared actual and planned data capture and DCS 2000 performance.

To determine the progress in modifying DCS 2000, we first reviewed plans and activities for system development, project management, risk management, testing, and deployment, and also reviewed their associated status and results reports. In addition, we attended contractor progress briefings presented to the bureau.

We interviewed bureau and contractor officials throughout our review. We performed our work at the Census Bureau's headquarters in Suitland, Maryland; the DCS 2000 program office in Lanham, Maryland; and the bureau's development contractor's facilities in Bowie, Maryland, from May 2000 through September 2000, in accordance with generally accepted government auditing standards.

We provided a draft copy of this letter to the Department of Commerce, which responded that it appreciated the opportunity to provide comments but that it had none.

First-Pass Data Capture Operations Have Progressed According to Plans

The bureau's master schedule for completing the 2000 decennial census specifies dates for the completion of significant activities, such as the first pass and second pass of data capture operations. In addition, the bureau has projected workloads and set performance goals for data capture operations, including contractual requirements for DCS 2000's performance.

The bureau's first-pass data capture operations have progressed according to plans, and DCS 2000's performance in support of first pass-operations has exceeded the bureau's goals. Since our May testimony, the bureau reported that it met the May 26 deadline for completing data capture from census questionnaires that were mailed

back. In addition, the bureau completed remaining first-pass data capture operations, such as following up with households that did not respond to a mailed questionnaire (nonresponse follow-up) ahead of schedule. As of August 23, the DCCs had received and checked in about 161 million census forms, as compared to about 152 million planned. At that time, the Baltimore, Pomona, and Phoenix DCCs had completed processing their first-pass form workloads.

On August 28, the bureau began second-pass processing at the Baltimore DCC, 2 weeks earlier than planned. Further, it began second-pass processing at the Pomona and Phoenix DCCs, 1 day earlier than planned, on September 11. On September 14, the Jeffersonville DCC completed first-pass processing, over 2 weeks ahead of schedule, and began the transition to second-pass processing. The bureau plans to initiate second-pass processing at the Jeffersonville DCC on October 10, 6 days ahead of schedule, and to conclude second-pass operations at all DCCs on November 15.

In supporting first-pass operations, DCS 2000 exceeded its accuracy and productivity goals. For example, the bureau reports that DCS 2000's OCR accuracy rate was about 99.3 percent at each DCC, exceeding the 98-percent accuracy goal. In addition, the bureau reports that the KFI accuracy rate was 97.6 percent or more at each DCC, exceeding the bureau's 96.5 accuracy goal, and the KFI keying rate exceeded the goal at each DCC.

DCS 2000 Second-Pass Modifications Are Progressing According to Schedule, and Testing Is Being Effectively Managed

In addition to the bureau's master schedule for completing the decennial census, the bureau and the DCS 2000 development contractor have a detailed schedule for modifying and testing DCS 2000 to support second-pass data capture operations. As of August 25, efforts to modify and test DCS 2000 were ahead of schedule. For example, operational testing at the Baltimore DCC was completed early, and second-pass operations began at the Baltimore DCC on August 28 (2 weeks early).

The modifications to DCS 2000 software do not involve a large of number of lines of code. However, the modifications are pervasive, affecting most system modules, and thus require extensive and effectively managed testing to ensure that the changes perform as intended and do not have unintended effects on the unmodified code.[6] To be effective, testing should include the testing of individual software units or modules, the integration of these units or modules in creating an application, and the integration of related applications in creating a system. Further, an operational test

[6]The second-pass changes require changes to fewer than 1,000 source lines of code, which is about 1.2 percent of the approximately 85,000 source lines of code in DCS 2000.

should be conducted to demonstrate that a system performs as intended when operated on-site by those expected to use it.

As we first reported in our May testimony, the contractor has taken an incremental approach to testing and has employed disciplined test-management processes that are consistent with best practices. As of August 25, the development contractor had

completed making second-pass-related modifications to DCS 2000 and had conducted software unit and integration and system integration testing, witnessed by bureau officials, to demonstrate that the system meets specified functional and performance requirements. The contractor's formal test report states that the system met all specified requirements. Further, operational testing at the Baltimore DCC also demonstrated that the modified DCS 2000 performs as intended. Although the bureau and its contractors have not yet prepared a formal test report for the Baltimore DCC second-pass operational test, they provided us with test results showing that the

system exceeded expected performance during the operational test and that all test objectives were met. For example, the test demonstrated DCS 2000's ability to retrieve and optically read 218,165 forms per day, versus the expected rate of 149,715 forms per day. Also, keyers were able to achieve 7,513 key strokes per hour versus the modeled rate of 4,400 key strokes per hour. The test also successfully demonstrated the transmission of census data to bureau headquarters. Further, DCS 2000's accuracy during the Baltimore test was validated by an independent organization.[7] This organization determined that (1) the OCR accuracy was 99.7 percent, exceeding the bureau's 98-percent goal, and (2) the keying accuracy rate was 98 percent, exceeding the bureau's 96.5-percent goal.

In addition, the DCS 2000 development contractor has continued taking steps to identify and mitigate risks. For example, because the integrity of the mass storage units that contain census long-form images is critical to the success of the two-pass approach, the contractor has taken steps to ensure that each form image is accounted for and available for second-pass operation. To accomplish this, the contractor made a tape back-up of the mass storage units, verified the tape back-up, and audited the mass storage units to determine whether each form image was stored as appropriate. As of August 23, the contractor's audit results showed that all of the more than 23.5 million images were accounted for on the mass storage devices.

Conclusions

The bureau has made significant progress toward completing first-pass data capture operations as planned, and during these operations DCS 2000 has performed as intended. Similarly, the bureau's development contractor has made significant

[7]RIT Research Corporation, a subsidiary of Rochester Institute of Technology, independently validated the Baltimore DCC operational test results.

progress toward modifying DCS 2000 to support second-pass data capture operations and has taken effective testing and risk management steps to ensure that the modified version of DCS 2000 is effectively implemented and performs as intended.

-- -- --

We are sending copies of this letter to the Honorable Norman Y. Mineta, Secretary of Commerce; the Honorable Kenneth Prewitt, Director of the U.S. Census Bureau; the Honorable Jacob J. Lew, Director of the Office of Management and Budget; and other interested parties. Copies will be made available to others upon request.

If you have any questions on matters discussed in this letter, please contact me at (202) 512-6240. Other key contributors to this letter include Mark Bird, Garry Durfey, and Richard Hung.

Randolph C. Hite
Associate Director
Governmentwide and Defense Information Systems

(512000)